Semantic Grouping of Call for Papers in Wikis

Sulaiman Alrayee
School of Information Studies
Al Imam University, Riyadh
Saudi Arabia
reyaee@gmail.com

ABSTRACT: *Weaving of semantic web has to go a long way. It depends on how build comprehensive as well as unified methods to group all related data. Now web data availability is increasing which leads to the creation of one system. In this paper we have introduced the concept of semantic grouping of wiki information. However, we need to long way to find semantic grouping.*

Keywords: Web 2.0, Semantic Web, Wiki, Call of papers

Received: 11 January 2012, Revised 2 March 2012, Accepted 8 March 2012

1. Introduction

For many reasons in information searching, it is useful to classify the semantic types into a smaller number of semantic groups. In an early work, the authors have established ûfteen high-level semantic groups that help reduce the conceptual complexity of the large domain covered by the UMLS [1]. In another related work the authors have used a diûerent attempt to partition the UMLS semantic network). Groupings of semantic types - the semantic groups - may prove to be useful in a number of applications including improved visualization and display of the knowledge in a particular domain [2]; natural language processing, where higher level categories are sometimes suûcient for semantic processing [3]; and auditing a domain for the valid representation of concepts and their interrelationships [4]. For example, if a particular concept in the semantic file such as thesaurus has multiple semantic types and this assignment leads to the concept appearing in two diûerent high-level groups, then it is possible that at least one of the semantic type assignments is incorrect. In our earlier work, the authors subjected the entire set of concepts in the 2000 version of the UMLS to this test, and we found a number of semantic type assignment errors through this method.

In this paper we concentrated on the main stream of Call for Papers as published in the wiki cfp that is used now as semantic store to reflect natural grouping as they offer certain advantages in their respective usage. On the one hand Web 2.0 provides more variety of authoritative data in more unstructured and machine un-friendly way which need certain heuristics and machine learning algorithms for knowledge exploration, on the other hand Semantic Web techniques provide limited datasets but in more structured form, which can easily be located and disambiguated. While we discussed this possibility in this paper by considering the facts in account, the aim of the paper is to develop an approach which encompasses Web 2.0 and Semantic Web technologies together to locate and aggregate person's relevant information into one user profile. We want to show that combining search engines along with intelligent use of semantic technologies and datasets related information can be located, disambiguated and

delivered to the user. Further we propose an application where the information found can be aggregated and presented in a coherent way as well as proving that Semantic web technologies and conventional web applications assist each other in better information management.

2. Related Work

Important information about a company or a person are nowadays not only stored on a traditional Web page but also in the Web 2.0 services like blogs, forums, Facebook, etc. Because the bulk of information is found online, having consolidated results is of great importance for both companies and people. While traditional search approaches like googling the name return a bulk of unsorted information, search services specialized for searching people or company data like Zoominfo [5], Pipl [6], and Intelius [7] are getting increasingly popular today.

There are really two separate groups of people that ZoomInfo is primarily aimed at, and that would be searchers who are looking either for themselves or various people in their lives (friends, family, colleagues, etc.), or human resources type folks who are looking to recruit their next employee. It will take a look at these two disparate groups separately.

Finding People Using ZoomInfo to find people is simple - just navigate to the home page and type in a name. Similarly, we can also look up company information by clicking on the "*Company*" text link above the main search query bar.

My profile had a couple of different links on there; mostly to blogs and Web sites that have cited in some way. We found the chances of editing our profile; this is where we could add a lot more to my ZoomInfo profile. This finally leads to search in popular search engines such as Google.

ZoomInfo offers Power Search as a great way to find employees - and according to their information, a lot of companies do use it: "(*our customers*) *include over 20% of the Fortune 500, 9 of the top 10 executive recruiting firms and thousands of others*." This is a paid service; however, I think that it would be well worth the money spent, especially for a top corporate firm looking to add a very specific area of expertise to their employee talent base.

We in our work recommend developing a type of semantic search system such as zoominfo which can enable us to develop semantic classification of wiki content. The wiki content we like to develop is for the wiki call for papers.

3. Datasets

In this paper we concentrate on one group of conference call for papers in different domains. Motivated from these related systems in Web 2.0 and Semantic Web domains we planned to experiment in our test application with the test set of call for papers on similar grounds. Our objective of this study is to combine both unstructured Web 2.0 information and structured Semantic Web information in one single system to find and present the information about member of wiki cfps by using Concept Aggregation Framework.

In the current work, we explain our approach using a data set of 297 categories of CFP consisting large number of individual listings. The number of CFPs during the access date was 23121. These CFPs are accessed by more than 100000 users every month. We have visited the CFPs and the wikicfp and understand the way of information organization in the pages. We did test crawling of the sample extracted CFPs and the crawling process has enabled us to search extensively and bring out the target notifications by using multilevel heuristic approach. The WIKICFP has enormous notifications specializing on different levels of categories across many domains.

The CFPs listed in the wiki page are now given below based on the number of CFPs listed according to the frequency. (Table 1)

Next the CFPs have several groups which remain unconnected as there is no broad classification. We in this work now proceed to group them into clusters based on the following semantic grouping as below.

The CFPs listed in wikicfp fall fewer than four categories. This means each of the CFP has four categories assigned by the organizers. Second, the CFPs have different main groups: Besides each CFP has a broad domain relation. Of those a high percentage of members of the humanities are less likely to use computers and the Web than members form the other groups.

Category	# CFPs	Category	# CFPs	Category	# CFPs
artificial intelligence	1136	computational biology	78	ontologies	40
communications	936	cognitive science	77	social web	40
software engineering	794	design	77	power engineering	39
security	776	natural language processing	77	electrical	38
data mining	705	healthcare	76	law	37
NLP	667	e-commerce	76	lasers	37
networking	642	measurement	76	social science	37
computer science	621	cryptography	76	model-driven development	37
databases	582	mobile computing	76	medical imaging	36
machine learning	516	life sciences	75	information system	36
multimedia	459	web 2.0	75	virtualization	36
signal processing	452	biotechnology	74	genealogy	36
engineering	406	soft computing	71	WORKSHOP	35
wireless	404	cloud	70	information theory	35
information retrieval	395	design automation	69	dependability	35
web	388	neural networks	69	cultural studies	35
image processing	376	chemistry	67	open source	35
robotics	355	industrial electronics	67	family history	35
HCI	352	power	67	computers	34
linguistics	333	literature	66	performance	34
education	324	database	66	programming	34
simulation	324	wireless communications	65	P2P	34
bioinformatics	321	collaboration	64	policy	32
computer	298	reliability	64	technologies	32
information technology	295	evolutionary computation	64	arts	32
software	295	web services	64	electro-optics	32
control	292	parallel processing	63	instrumentation	32
semantic web	282	biometrics	63	cyber-physical systems	32
modeling	277	games	62	environmental engineering	32
computer vision	254	data management	62	safety	31
embedded systems	232	medicine	61	CSCW	31
technology	229	computation theory	61	semiconductor	31
systems	219	theoretical computer science	61	human computer interaction	31
cloud computing	218	wireless sensor networks	61	computer security	31
communication	216	biomedical	60	business intelligence	31
information systems	214	ambient intelligence	59	theory	30
management	212	psychology	58	social	30
computer graphics	205	aerospace	58	agriculture	30
automation	200	microwave	57	RFID	30
distributed systems	197	services	56	social computing	30
networks	190	computer networks	56	recommender systems	30
computational intelligence	190	electrical engineering	56	linked data	30
pattern recognition	174	mechatronics	56	mechanical	29
mobile	174	middleware	56	graphics	29
privacy	172	history	55	AI	29
sensor networks	169	manufacturing	55	digital libraries	29
circuits	163	speech	54	evaluation	28
social networks	163	informatics	54	industrial	28
internet	160	real-time	54	optical	28
computer architecture	160	materials	53	XML	28
electronics	151	compilers	53	adaptation	28
e-learning	151	cybernetics	53	network management	28
programming languages	151	power electronics	53	augmented reality	28
distributed computing	148	antennas	53	software architecture	28
information	141	VLSI	53	language	27

electron devices	141	knowledge engineering	53	interaction	27
parallel computing	136	e-business	53	society	27
business	135	marketing	52	navigation	27
algorithms	131	operating systems	51	radar	27
economics	127	interdisciplinary	51	logistics	27
ubiquitous computing	126	learning	50	autonomic computing	27
computing	122	knowledge representation	50	software testing	27
environment	120	innovation	49	systems engineering	26
energy	118	research	48	remote sensing	26
pervasive computing	114	SYSTEM	47	neuroscience	25
knowledge management	114	sustainability	47	machine translation	25
applications	112	ICT	47	parallel programming	25
intelligent systems	110	virtual reality	47	english	24
network	106	grid computing	47	ethics	24
science	105	culture	46	religion	24
formal methods	105	information management	46	IT	24
agents	101	finance	45	intelligent transportation	24
biomedical engineering	101	grid	45	music	24
information security	96	sociology	44	USENIX	24
optimization	93	propagation	44	peer-to-peer	24
verification	93	art	44	globalization	23
network security	92	photonics	44	usability	23
semantics	90	ehealth	44	sensor	23
biology	90	multi-agent systems	44	complexity	23
mathematics	90	internet of things	44	web mining	23
visualization	90	information science	43	requirements engineering	23
knowledge discovery	90	politics	43	wireless network	23
modeling	89	SOA	43	development	22
architecture	88	elearning	43	applied computing	22
logic	86	physics	42	video	22
mobility	85	GIS	42	fault tolerance	22
social media	85	telecommunication	42	health informatics	22
social sciences	85	ubiquitous	42	business management	22
high performance computing	83	molecular biology	42	IFAC	21
computational linguistics	81	civil engineering	42	embedded	21
humanities	81	statistics	41	optoelectronics	21
health	81	text mining	41	computational science	21
sensors	81	human-computer interaction	41	wireless communication	21
wireless networks	80	mechanical engineering	41	e-health	21
philosophy	79	testing	40	collective intelligence	21
telecommunications	79	IR	40	complex systems	21
trust	79	pervasive	40	languages	20
ontology	78	computer engineering	40	multidisciplinary	20
nanotechnology	78	industry applications	40	media	20

Table 1. Wiki CFP Subject classification with frequency of cfps

Third, CFPs come from over different subject bases and different countries. Information is therefore found in many different languages. Multilanguage websites can pose a problem for semi-automatic detection of information about the conference. Ideally the heuristics form extracting the right information should be done in many different languages, and after detecting the language of the website some translation software or some other heuristics should be used.

In order to group the CFP by natural semantic categories, we took the categories of the CFPs assigned by the authors who produced CFPs. The CFPs and their subthemes are now classified based on the proposition, called Group alignment algorithm as below.

The aim of this exercise is to group the categories of CFPs based summation algorithm which is clear and direct. To accumulate

a group of m groups, instead of (m -1) standard additional categories, where all exponent comparisons and category shifting are scattered, we collect them together to form a unique exponent comparator in each step. Now, all categories in each group are aligned into consistent fixed-point format, so can be summed up with simple fixed-point accumulator in the next step. The Group-Alignment based Accurate Floating-Point Summation was earlier successfully deployed by Chuan et al [8]. Figure 1 reveals now the steps of this floating-point summation algorithm. This semantic grouping works well and the resulting applied groups of the selected CFPs are given in the figure 2.

A set of n categories $s_1, s_2, \ldots, s_n$, which are partitioned into groups with m sub groups

Output: The summation result S

For k =1 to n / m

1. *Split each sub group s_i in this group into two parts as Fraction F_i and Exponent E_i, Restore the hiding category in F_i to form the group M_i*
2. *Find the the largest Exponent E_{max} within E_i*
3. *Calculate every $(E_{max} - E_i)$*
4. *Shift M_i to right by $(E_{max} - E_i)$ bits, round the shifted category to nearest-even if necessary.*
5. *Sum up all shifted M_i*
6. *Feed rounded M_{sum} and E_{max} to a simplified group formatter*

End For

Figure 1. Group-alignment Algorithm

In the figure 2, the letters denote the categories. The arrows and other connections specify how the categories come together in semantic groupings. Thus, looking at the situation in hand sight it is clear that our techniques would produce still better results for e.g. a group of CFPs in natural sciences using a common language. However, this does make our modest success even more significant.

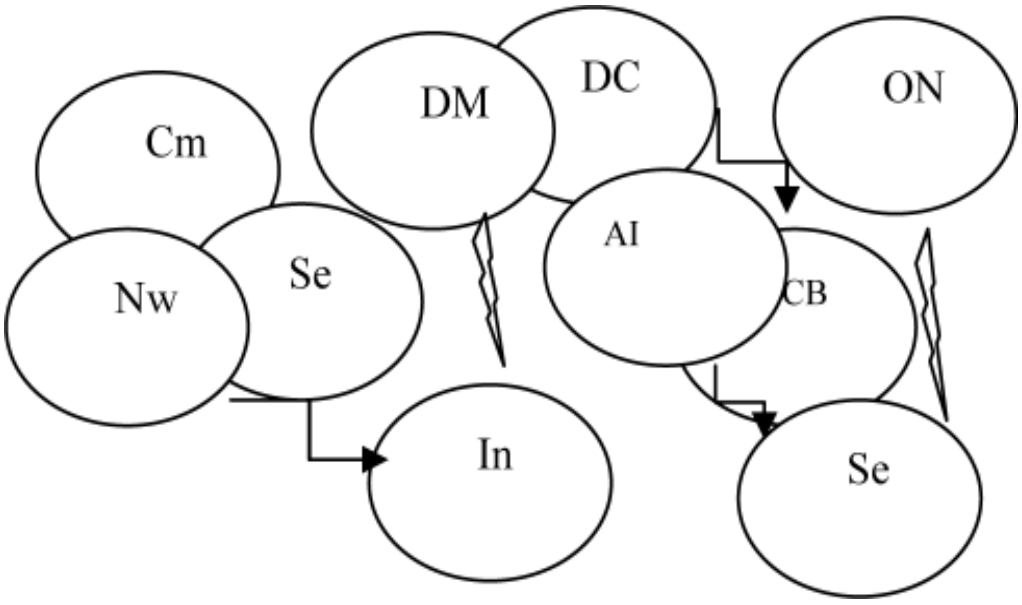

Figure 2. Semantic grouping based on floating point algorithm

4. Future work

This work reports the first level work we did on grouping call for papers in a semantic way. We plan to introduce multidimensional clustering that can enable us to prepare clusters of related conferences. Further we plan to have a natural grouping not based on the classification given by the conference managers, but based on the content. The content of the call for papers will be subjected to natural language parsing and will lead to the natural cluster formation.

References

[1] McCray, AT., Burgun, A., Bodenreider, O. (2001). Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo,10 (Pt 1) 216–20.

[2] Nelson SJ., Sheretz, DD., Tuttle MS, Erlbaum MS. (1990). Using MetaCard: a HyperCard browser for biomedical knowledge sources, Proc Annu Symp Comput Appl Med Care,151–4.

[3] Rindûesch, TC., Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform, 36 462–77

[4] Cimino, JJ. (1998). Auditing the uniûed medical language system with semantic methods. J Am Med Inform Assoc, 5(1) 41–51.

[5] Inaba, A., Supnithi, T., Ikeda, M., Mizorguchi, R., & Toyoda, J. (2000). How Can We Form Effective Collaborative Learning Groups. *In*: Proceedings of the 5th International Conference on Intelligent Tutoring Systems, 282-291.

[6] Kumar, V. (1992). Algorithms for Constraint Satisfaction Problems: A Survey. *AI Magazine*, 13 (1) 32-44.

[7] Leone, N., Pfeifer, G., Faber, W., Deiter, T., Gottlob, G., Perr, S., & Scarcello, F. (2006). The DLV system for knowledge representation and reasoning. *ACM Transactions on Computational Logic*, 7 (3) 499–562.

[8] He, Chuan., Qin, Guan., Lu, Mi ., Zhao, Wei (2010).Group-Alignment based Accurate Floating-Point Summation on FPGAs. citeseerx.ist.psu.edu/